Credit EDA Assignment

SUBMITTED BY,

SINDHU L

DS C63 BATCH

TABLE OF CONTENTS

PROBLEM STATEMENT

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. So here in this study we will understand the driving factors behind loan default and the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
- Using Exploratory Data Analysis we are going to analyze the pattern present in bank data which helps the company to reduce their risk associated such as :

*If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

*If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

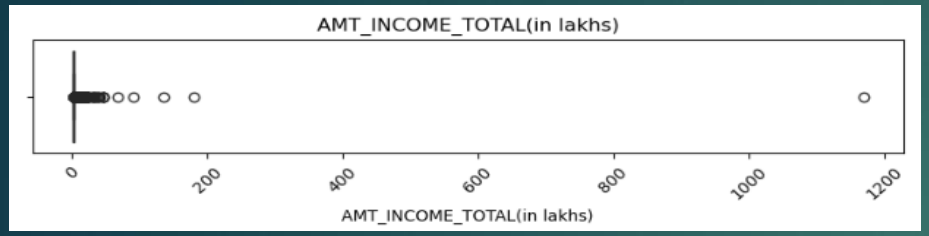
APPROACH

1. Data understanding
2. Data Cleaning
   * Data Quality Check
   * Handling missing values and Imputation
   * Handling Outliers
3. Univariate Analysis
   * Categorical Ordered Analysis
   * Categorical Unordered Analysis
   * Numerical Analysis
4. Data Imbalance check
5. Univariate Segmented Analysis
6. Bivariate Analysis
   *Numerical-Numerical Analysis
   * Numerical -Categorical Analysis
   * Categorical - Categorical Analysis
7. Multivariate Analysis
8. Analysis on Previous Application
9. Merged Data Analysis

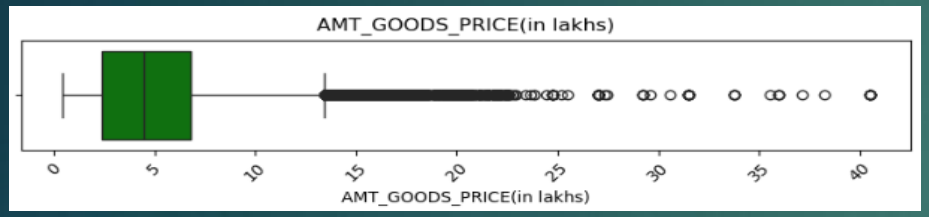
ANALYSIS OF CURRENT APPLICATION DATA

- For this analysis we were provided with application_data.csv file
- This contains the information regarding the current application
- It has data whether the Client has payment difficulties or not

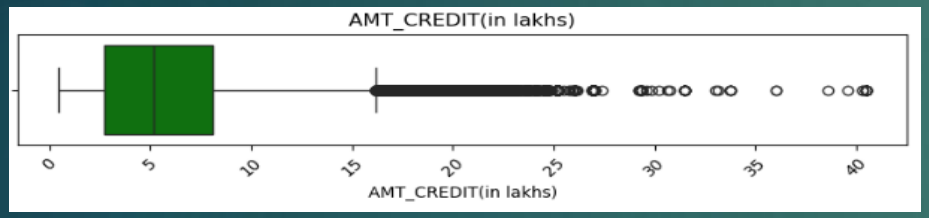
IDENTIFYING OUTLIERS



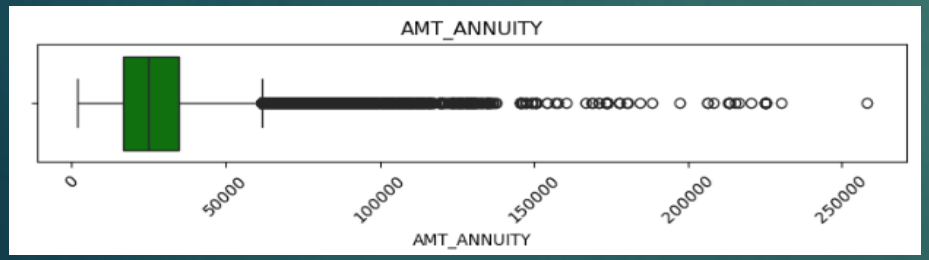
- Total Income have significantly higher outlier which implies clients with very much higher incomes are included in analysis



- Goods Price data looks to have continuous values till certain range; It still has outliers on higher range



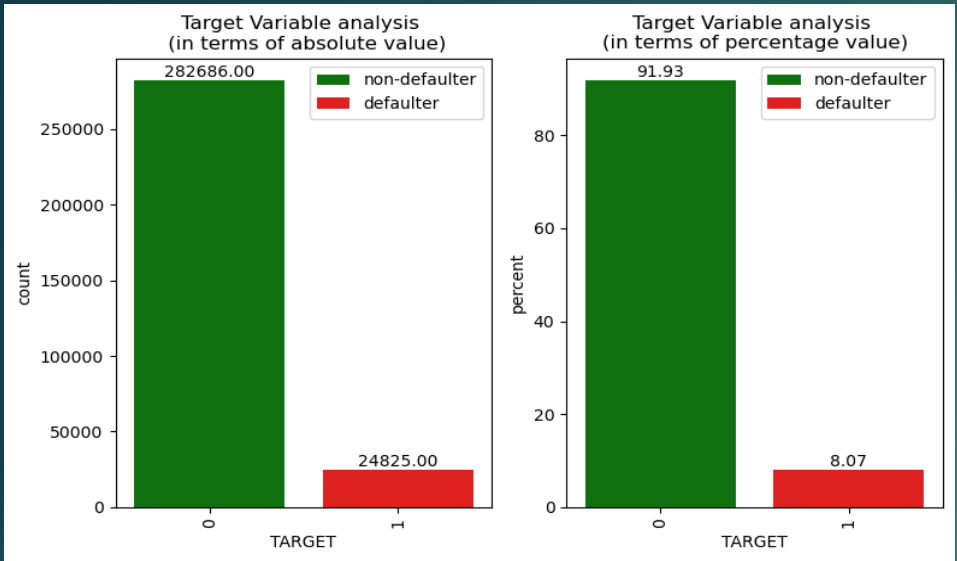
- Amount credited has outliers in them which implies higher range of amounts have been credited to certain group of people



- Outliers present in Amount Annuity tells that there are group of clients who pay higher annuity amount

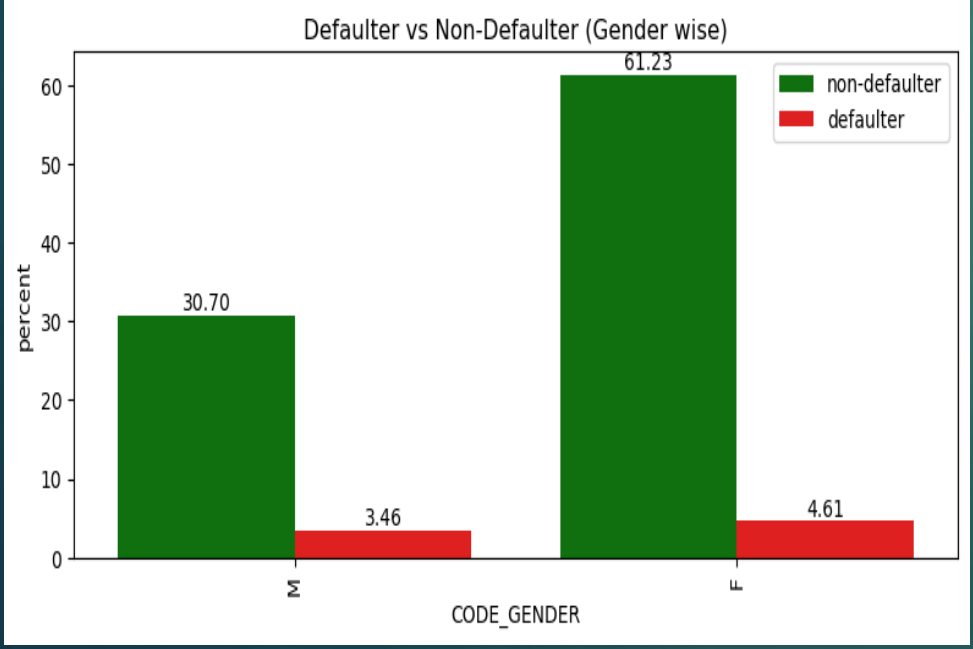
- No of Days employed have significantly higher outlier; This might be due to typo error or human error because such values cant exist

RESULTS OF CURRENT APPLICATION DATA



DATA IMBALANCE IN TARGET VARIABLE

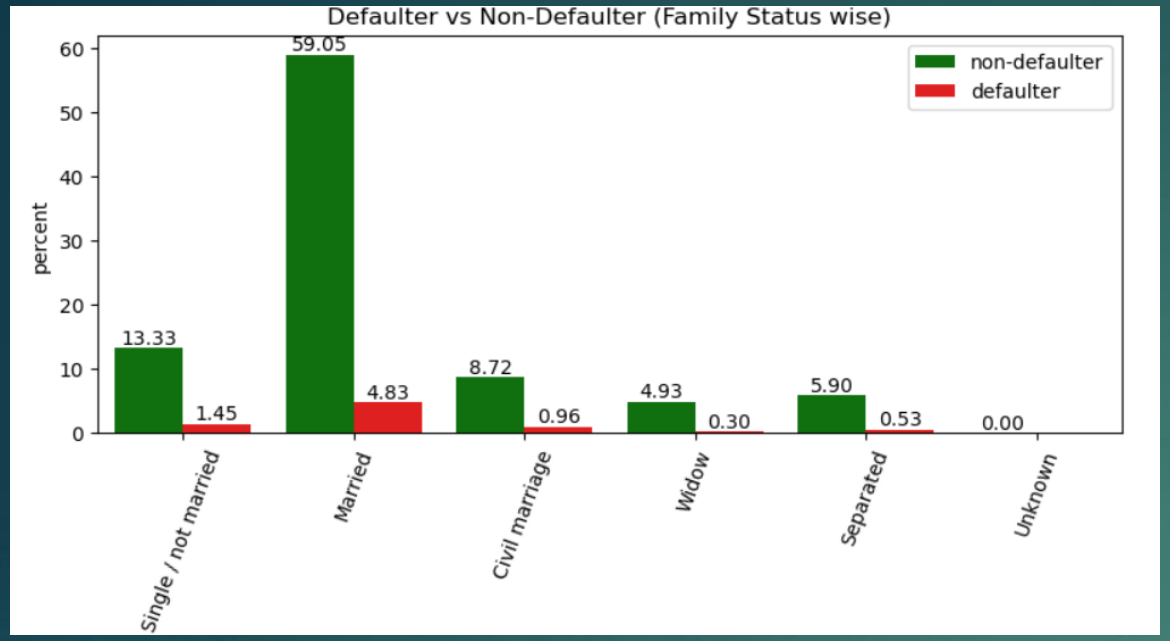
- Non-defaulter:92%, Defaulter:8%
- 11 times Non-Defaulter category dominates
- 1 in 11 person will be a default
- This imbalance is good for company as people who pays the amount correctly are more; financial loss would be less

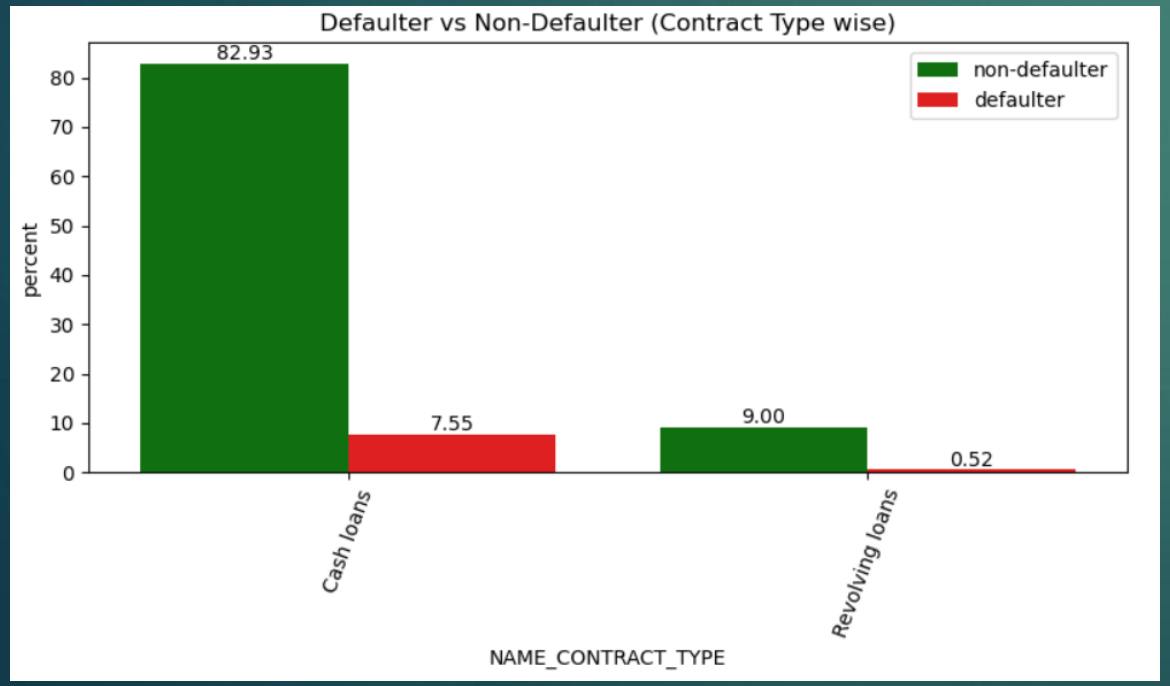


CLASSIFICATION BETWEEN DEFAULTERS AND NON-DEFAULTERS

GENDER WISE

- More number of Female candidates have applied for loans than Male candidates
- Male candidates are defaulting more than Female candidate by seeing ratio between defaulters and non-defaulters
- Default percent
  - *Male-10%
  - *Female-7%





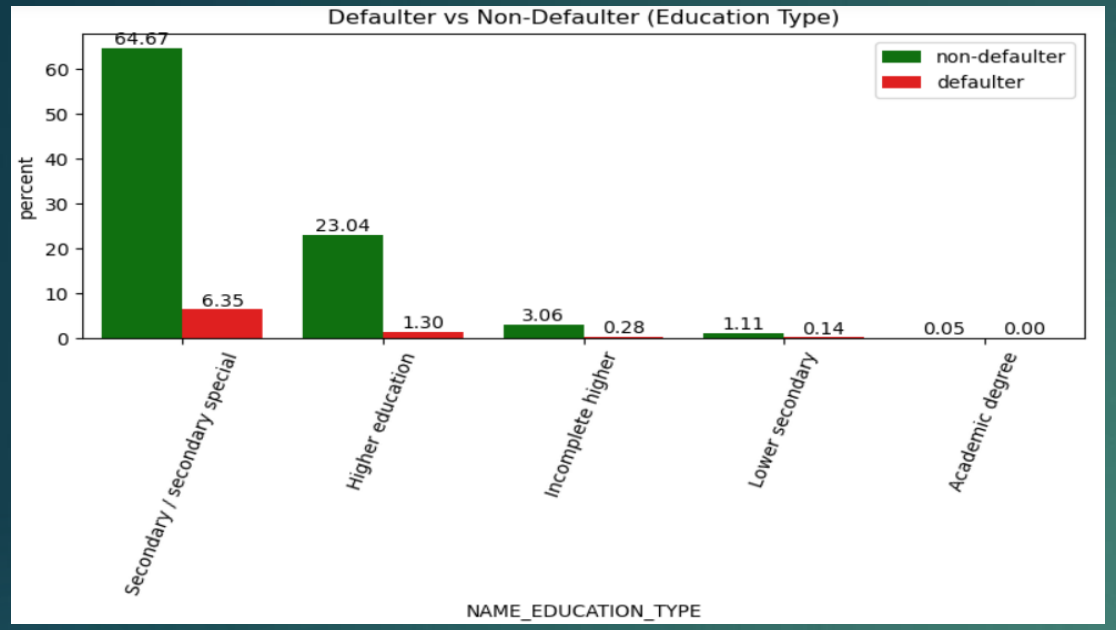
CLASSIFICATION BETWEEN DEFAULTERS AND NON-DEFAULTERS

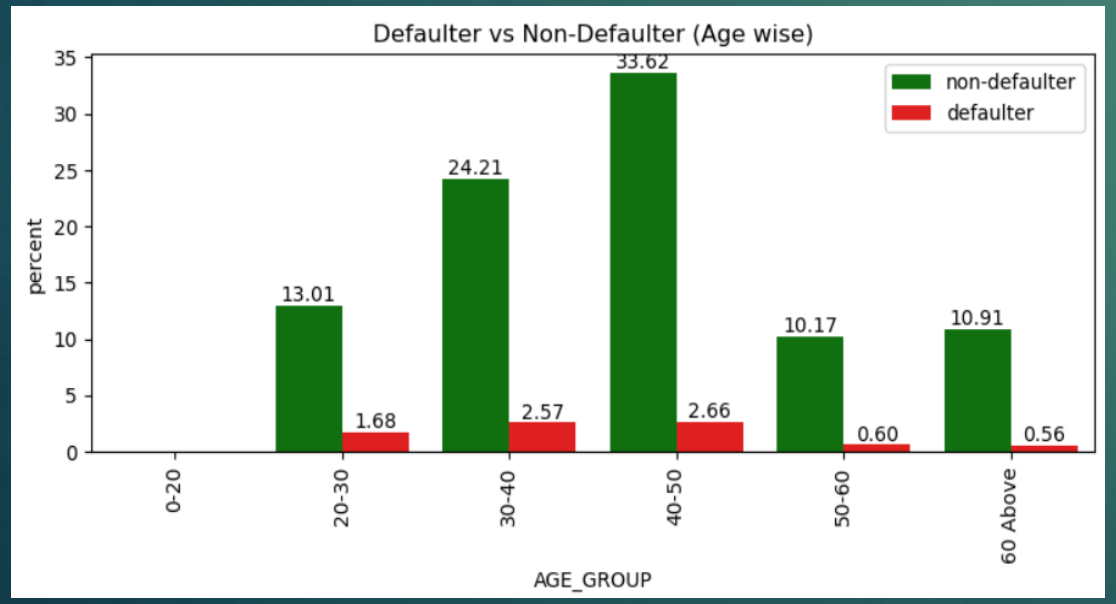
FAMILY STATUS WISE

- More number of clients who have applied loans are Married
- By calculating ratio between defaulters and non-defaulters of each category we can see,
  *Less Risk Category: Widow (But their numbers are less)
  *More Risk Category: Single/not married, civil marriage

CONTRACT TYPE WISE

- Cash loans have been given in larger numbers than revolving loans
- By calculating ratio between defaulters and non-defaulters of each category we can see,
  *Less Risk Category: Revolving loans
  *More Risk Category: Cash loans





CLASSIFICATION BETWEEN DEFAULTERS AND NON-DEFAULTERS

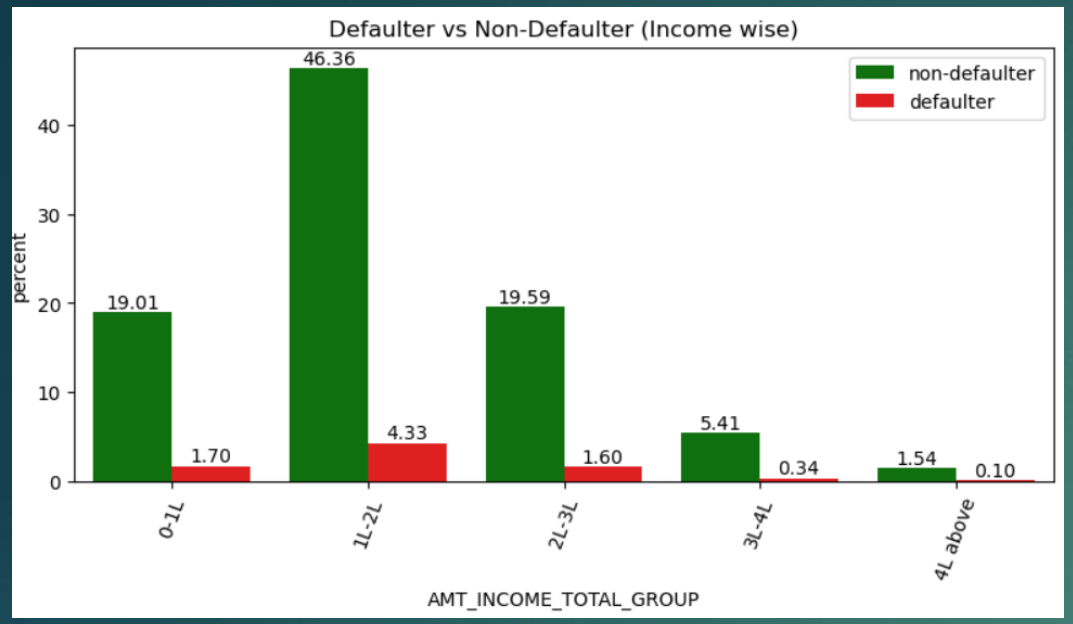
EDUCATION TYPE WISE

- Clients with Secondary education are in larger numbers
- By calculating ratio between defaulters and non-defaulters of each category we can see,
  *Less Risk Category: Higher education, Academic degree
  *More Risk Category: Lower secondary

AGE WISE

- More clients between 40-50 have applied loans
- By calculating ratio between defaulters and non-defaulters of each category we can see,
  *Less Risk Category: 60 above,50-60(People with higher age)
  *More Risk Category: 20-30,30-40 (People with less age)

CLASSIFICATION BETWEEN DEFAULTERS AND NON-DEFAULTERS

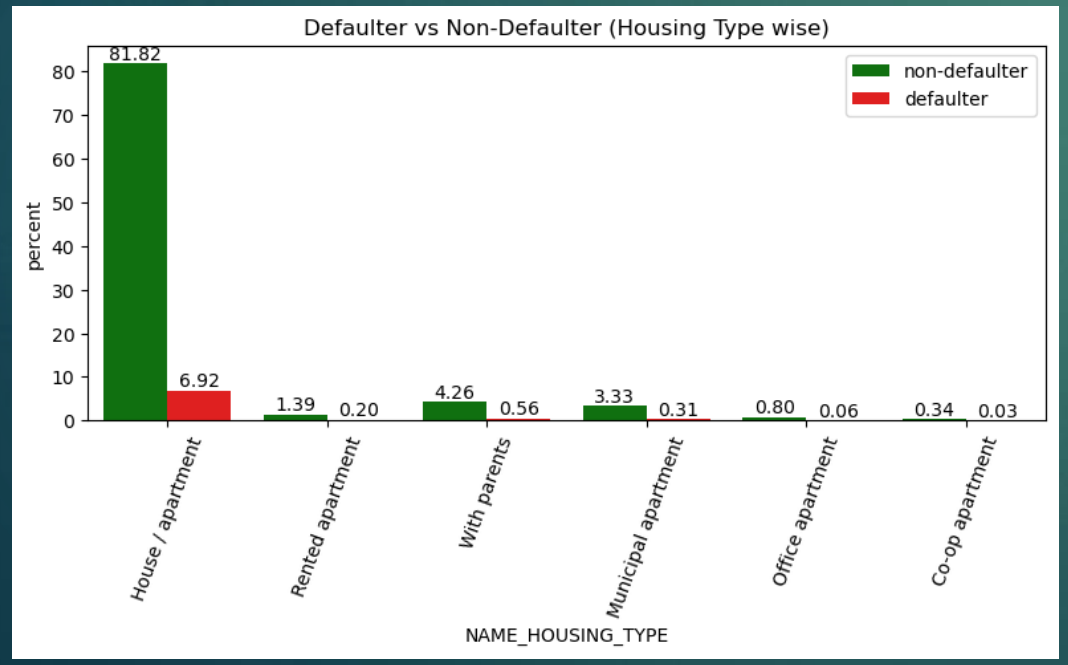


INCOME WISE

- Clients with Income range between 1-2Lakhs are in larger numbers
- By calculating ratio between defaulters and non-defaulters of each category we can see,

  *Less Risk Category: 3-4L,4L above (People with Higher income)

  *More Risk Category: 1-2L followed by 0-1L (People with lesser income)

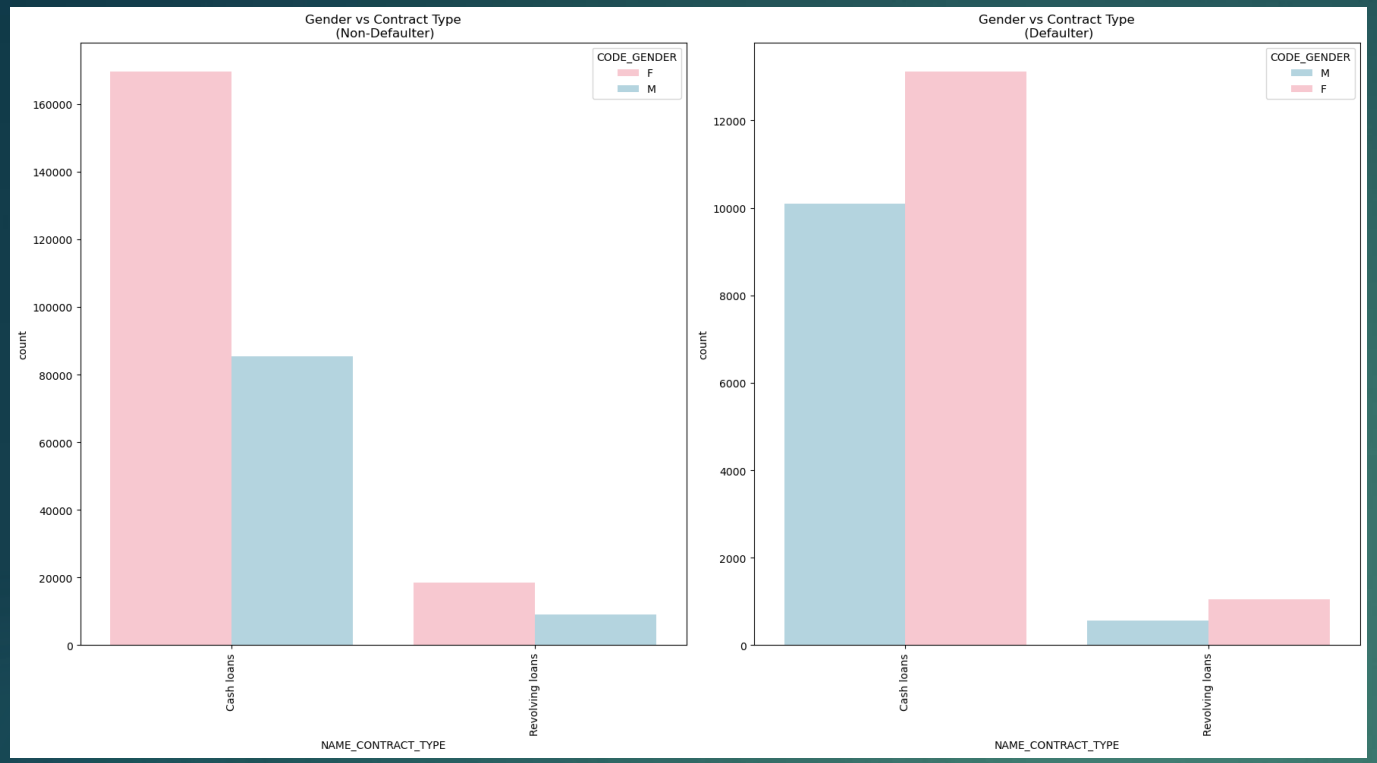


HOUSING TYPE WISE

- Clients with House/Apartments are in larger numbers
- By calculating ratio between defaulters and non-defaulters of each category we can see,

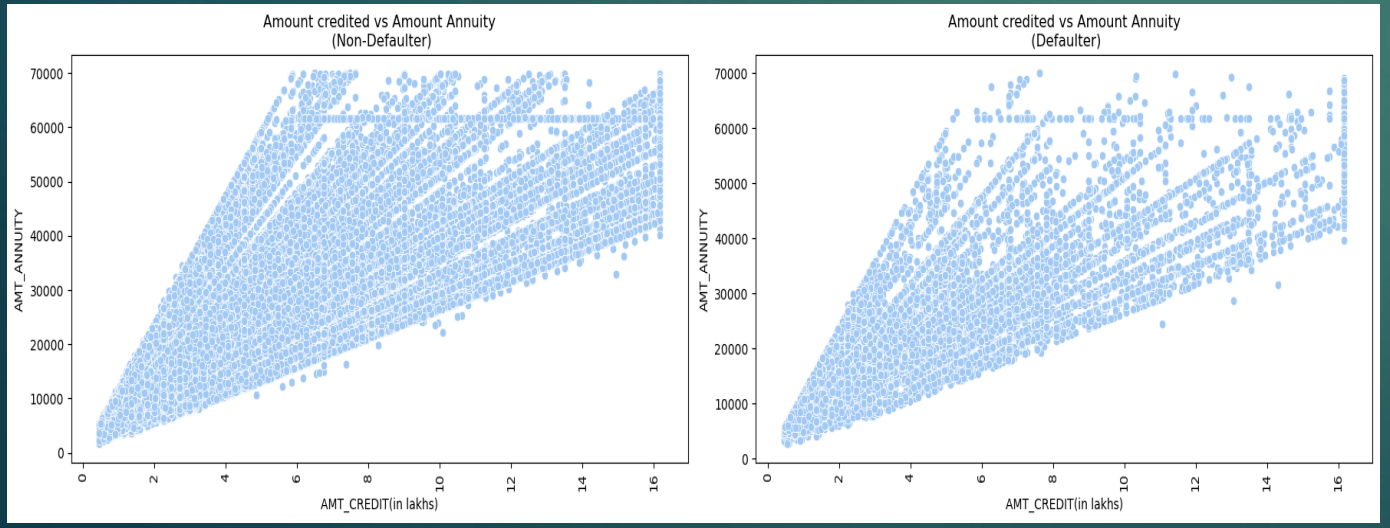
  *Less Risk Category: Office Apartment, House/Apartment

  *More Risk Category: Rented Apartment, With Parents



CONTRACT TYPE VS GENDER

- Female candidates are predominantly seen in both contract types as they have applied more number of applications
- Male candidates are defaulting more

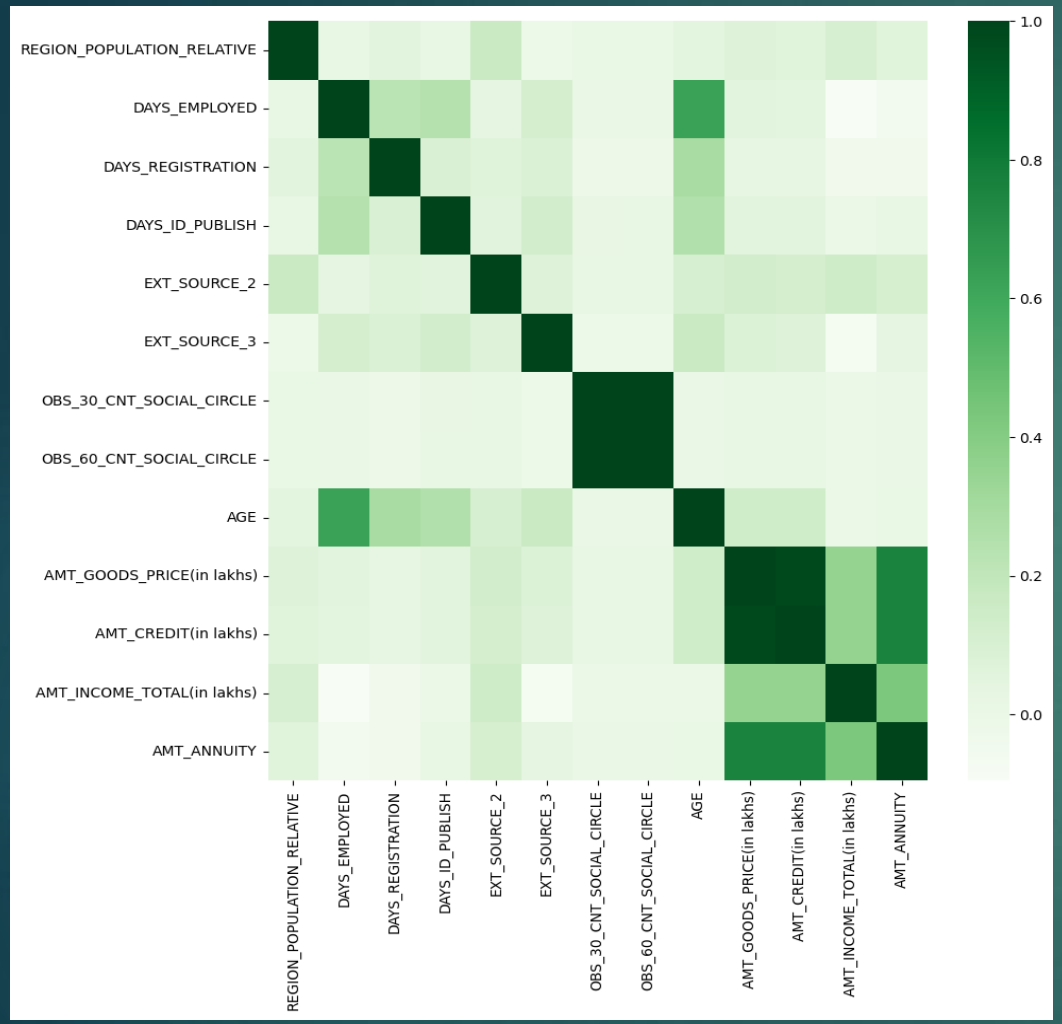


AMOUNT CREDITED VS AMOUNT ANNUITY

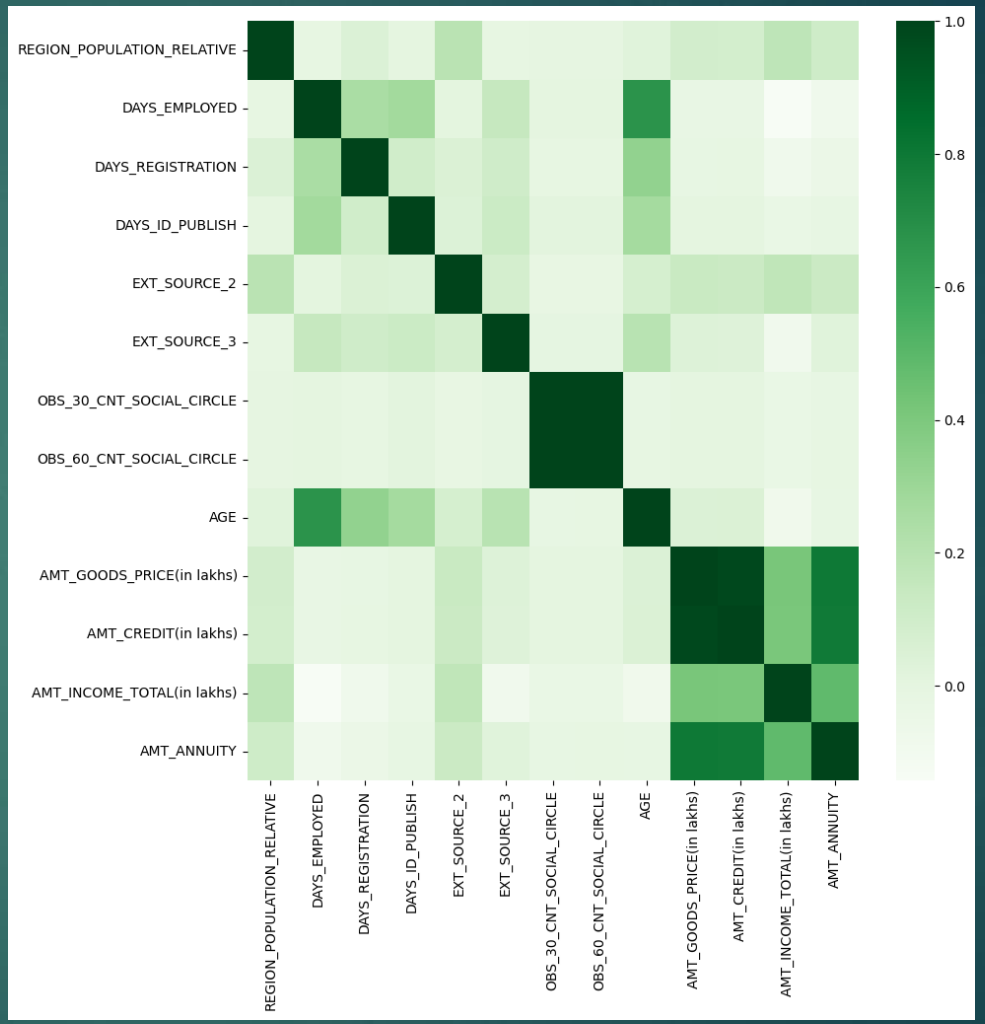
- There is a good correlation seen between them
- They are directly proportional

CORRELATIONS OF VARIABLES IN DEFAULTERS AND NON-DEFAULTERS

DEFAULTER



NON-DEFAULTER



*Both the plots looks same in terms of correlations

*Strong correlations are seen between Goods price, Amount credited ,Amount annuity

*Strong correlations are seen between OBS_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE

TOP 10 CORRELATIONS OF VARIABLES IN DEFAULTERS AND NON-DEFAULTERS

DEFAULTER

	column_1	column_2	correlation
0	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.997877
1	AMT_GOODS_PRICE(in lakhs)	AMT_CREDIT(in lakhs)	0.982183
2	AMT_GOODS_PRICE(in lakhs)	AMT_ANNUITY	0.760224
3	AMT_CREDIT(in lakhs)	AMT_ANNUITY	0.759713
4	DAYS_EMPLOYED	AGE	0.626751
5	AMT_INCOME_TOTAL(in lakhs)	AMT_ANNUITY	0.430026
6	AMT_GOODS_PRICE(in lakhs)	AMT_INCOME_TOTAL(in lakhs)	0.352915
7	AMT_CREDIT(in lakhs)	AMT_INCOME_TOTAL(in lakhs)	0.351081
8	DAYS_REGISTRATION	AGE	0.289114
9	DAYS_ID_PUBLISH	AGE	0.252863

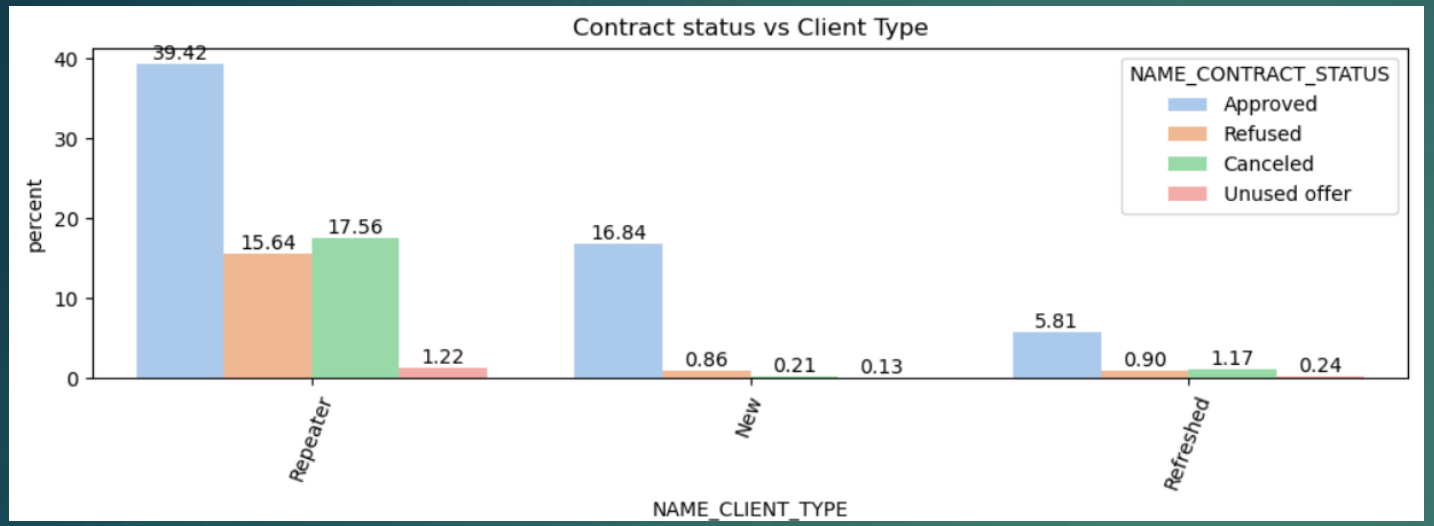
NON-DEFAULTER

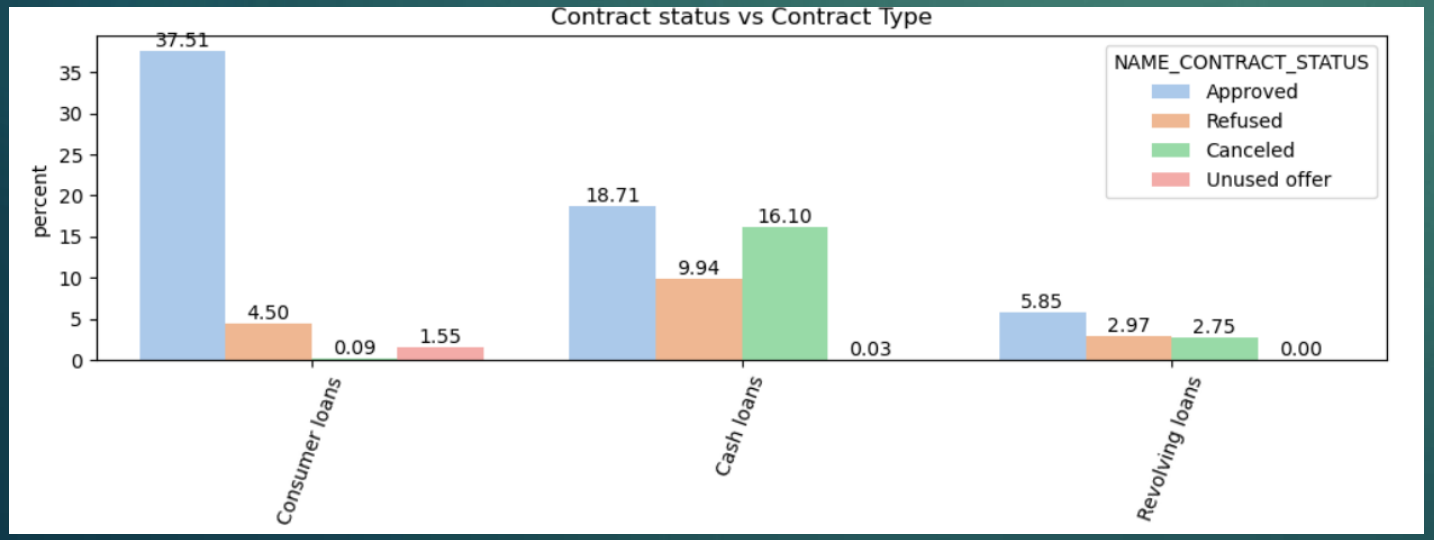
	column_1	column_2	correlation
0	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.997858
1	AMT_GOODS_PRICE(in lakhs)	AMT_CREDIT(in lakhs)	0.985841
2	AMT_GOODS_PRICE(in lakhs)	AMT_ANNUITY	0.795556
3	AMT_CREDIT(in lakhs)	AMT_ANNUITY	0.792639
4	DAYS_EMPLOYED	AGE	0.674929
5	AMT_INCOME_TOTAL(in lakhs)	AMT_ANNUITY	0.486171
6	AMT_GOODS_PRICE(in lakhs)	AMT_INCOME_TOTAL(in lakhs)	0.411538
7	AMT_CREDIT(in lakhs)	AMT_INCOME_TOTAL(in lakhs)	0.408056
8	DAYS_REGISTRATION	AGE	0.333151
9	DAYS_EMPLOYED	DAYS_ID_PUBLISH	0.280191

- Very strong correlation between OBS_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE in both cases
- Goods price and amount credited are very well correlated in both cases; suggests people purchasing costly goods get higher amount credited
- Most of the correlations are same in both cases
- Correlation values are slightly lower in defaulter side

RESULTS OF PREVIOUS APPLICATION DATA

- previous_application.csv data set contains information regarding the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**





CONTRACT STATUS VS CLIENT TYPE

- Clients who are repeaters have applied more number of previous loan application
- Most of the loans are approved for repeaters; But Repeaters loans were also cancelled in large numbers

CONTRACT STATUS VS CONTRACT TYPE

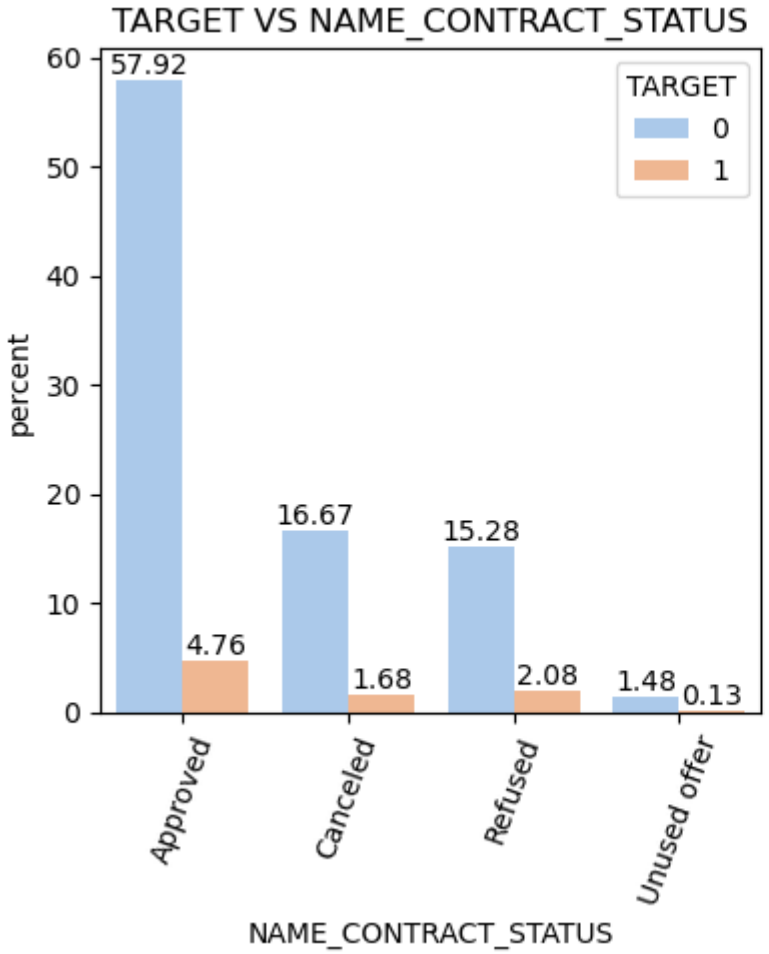
- Consumer loans have been issued in large numbers
- Consumer loans have been approved mostly; But clients are refusing and leaving them unused in more numbers comparatively

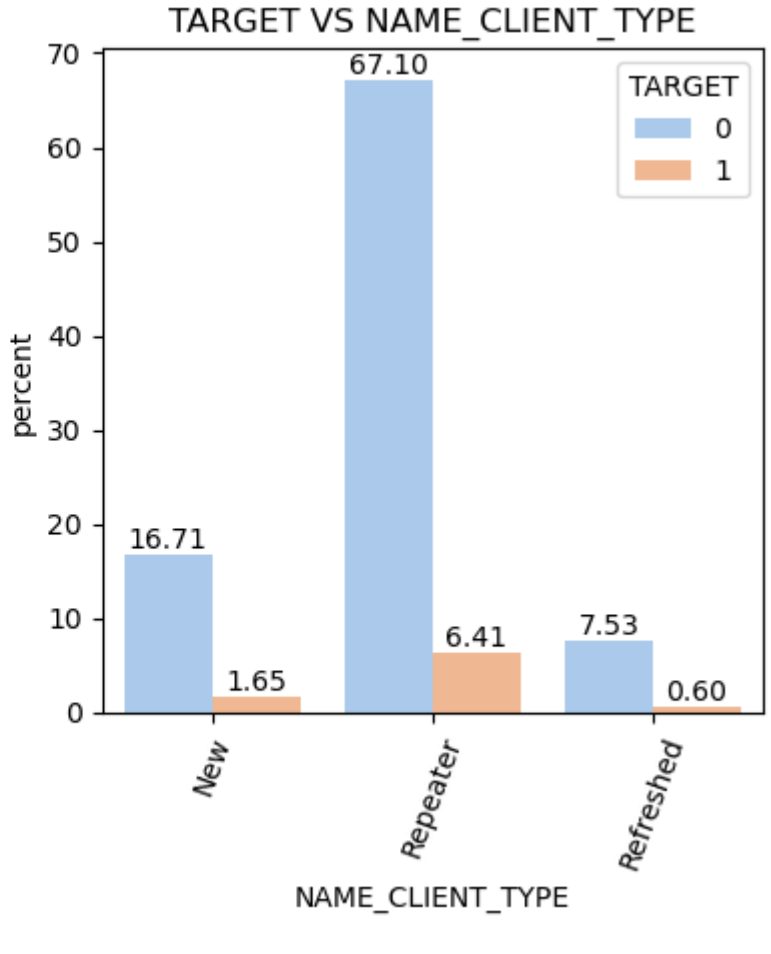
CORRELATIONS OF VARIABLES IN PREVIOUS APPLICATION

	column_1	column_2	correlation
0	AMT_APPLICATION(in lakhs)	AMT_GOODS_PRICE(in lakhs)	0.993430
1	AMT_CREDIT(in lakhs)	AMT_GOODS_PRICE(in lakhs)	0.989196
2	AMT_APPLICATION(in lakhs)	AMT_CREDIT(in lakhs)	0.941220
3	AMT_ANNUITY	AMT_GOODS_PRICE(in lakhs)	0.871289
4	AMT_ANNUITY	AMT_CREDIT(in lakhs)	0.859164
5	AMT_ANNUITY	AMT_APPLICATION(in lakhs)	0.833208
6	CNT_PAYMENT	AMT_APPLICATION(in lakhs)	0.695709
7	CNT_PAYMENT	AMT_GOODS_PRICE(in lakhs)	0.690232
8	CNT_PAYMENT	AMT_CREDIT(in lakhs)	0.658099
9	AMT_ANNUITY	CNT_PAYMENT	0.455625

- Strong correlation among Goods price ,Application amount, Amount credited
- As seen in Current application data, strong relation among Annuity Amount, Amount credited, Application amount

MERGED DATASET RESULTS





CONTRACT STATUS VS TARGET

- Comparatively the Approved loans have lesser default
- This is a good sign for company ; but still defaulters exists in approved loans

CLIENT TYPE VS TARGET

- New Clients have more default than other groups

RECOMMENDATIONS

- The loan company should prefer more female candidates as they are less likely to default compared male candidates
- Issuing more revolving loans should be preferred, as the people who opt for this kind of loan repay better
- Clients with good education quality has to be preferred; Higher the education they repay better
- Clients with older age group are preferred as they are less likely to default; For younger age people extra cautious while approving loans is recommended
- Clients who are repeating are preferred

CONCLUSION

- Finally, we could get the insights from the data provided; By which we could categorize the people according to the risk criteria
- Using EDA process we identified the patterns and driving factors behind loan default
- This overall process enhanced our understanding of loan applicants characteristics and how to analyze them
- These results would help the loan providing company to enhance their business